

# Spatiotemporal Short-term Traffic Forecasting using the Network Weight Matrix and Systematic Detrending

Alireza Ermagun<sup>1</sup>, David Levinson<sup>2</sup>

<sup>1</sup> Postdoctoral Fellow, Northwestern University, Department of Civil and Environmental Engineering, Technological Institute, 2145 Sheridan Rd, Evanston, IL 602088 USA  
Email: [alireza.ermagun@northwestern.edu](mailto:alireza.ermagun@northwestern.edu) (*Corresponding Author*)

<sup>2</sup> Professor, University of Sydney, School of Civil Engineering  
Email: [david.levinson@sydney.edu.au](mailto:david.levinson@sydney.edu.au)

## ABSTRACT

This study examines the dependency between traffic links using a three-dimensional data detrending algorithm to build a network weight matrix in a real-world example. The network weight matrix reveals how links are spatially dependent in a complex network and detects the competitive and complementary nature of traffic links. We model the traffic flow of 140 traffic links in a sub-network of the Minneapolis - St. Paul highway system for both rush hour and non-rush hour time intervals, and validate the extracted network weight matrix. The results of the modeling indicate: (1) the spatial weight matrix is unstable over time-of-day, while the network weight matrix is robust in all cases and (2) the performance of the network weight matrix in non-rush hour traffic regimes is significantly better than rush hour traffic regimes. The results of the validation show the network weight matrix outperforms the traditional way of capturing spatial dependency between traffic links. Averaging over all traffic links and time, this superiority is about 13.2% in rush hour and 15.3% in non-rush hour, when only the 1<sup>st</sup>-order neighboring links are embedded in modeling. Aside from the superiority in forecasting, a remarkable capability of the network weight matrix is its stability and robustness over time, which is not observed in spatial weight matrix. In addition, this study proposes a naïve two-step algorithm to search and identify the best look-back time window for upstream links. We indicate the best look-back time window depends on the travel time between two study detectors, and it varies by time-of-day and traffic link.

**Keywords:** Traffic Forecasting; Spatial correlation; Competitive links; Traffic Flow; Weight matrix

# 1 Introduction

The growing interest in short-term traffic forecasting has been stimulated by the ubiquity of Intelligent Transportation Systems (ITS) technologies (Vlahogianni et al., 2005). Forecasting traffic conditions responds to an increasing need to provide accurate information to drivers, and consequently assisting them on selecting appropriate paths. Transport analysts tackled traffic forecasting, while focusing on the temporal dependency of traffic conditions in a road segment (Smith et al., 2002). This research stream forecasts the traffic condition of a given location for one or more discrete time intervals, using only historical data at the forecast location. However, the traffic condition of a road segment is not only related to its previous condition, but also to the condition of the adjacent road segment (Okutani and Stephanedes, 1984). In essence, vehicles, which will be at one location soon, are already in the network upstream of the relevant location, and of typical patterns of flow. This phenomenon opened a new gateway to the short-term traffic forecasting research stream, suggesting embedding the spatial component in modeling.

Including the spatial component in short-term traffic forecasting benefited from the traffic information of upstream feeder links to achieve better traffic flow prediction (Okutani and Stephanedes, 1984). Although this approach was the focal point in the literature of spatiotemporal short-term traffic forecasting, two unanswered questions remained:

- What links have a spatial correlation with the study link, and should their information be included in the model?
- How should the spatial information of selected links be captured?

Two techniques have evolved to deal with these questions (Ermagun and Levinson, 2016a). The *naïve* technique assumes the traffic condition of a study link is highly correlated with the upstream in uncongested regimes and with both upstream and downstream in congested regimes. It nominates links in which their traffic conditions are highly correlated with the study link as potential spatially correlated links (Sun et al., 2005).

In contrast, the  $l^{th}$ -order neighbors technique takes a step forward in more comprehensively measuring the spatial correlation between traffic links, assumes the spatial correlation is not limited to only the adjacent upstream and downstream. To capture the spatially correlated links, this technique employs either the correlation-coefficient or distance adjustment methods. It recommends different adjacent rings of neighboring links as potential spatially correlated links, while it speculates that the magnitude of correlation decreases by distance (Kamarianakis et al., 2004). For example, the first-order captures the dependency between the study link and its immediate adjacent links. If the adjacent links are confined to upstream and downstream, this approach collapses to the naïve technique.

While embedding the spatial component in forecasting methods acts as a catalyst, its functioning is hindered by the constraints of techniques in capturing spatially correlated links. The positivity of spatial correlations postulates that traffic links have a positive spatial dependency. In essence, this hypothesis represents complementary (upstream and downstream) traffic links. For simple single facility corridors, this may be sufficient. On the flip side of the coin is the competitive nature of traffic links. Competitive links bear a significant proportion of diverted vehicles, when one link is saturated or closed (Ermagun and Levinson, 2016b). Short-term traffic forecasting has historically been confined to examining complementary links. In consequence, the competitive

nature of traffic links has been overlooked in capturing spatially correlated links and short-term traffic forecasting.

This research extends the attempt to capture spatial correlation between traffic links, and adds to the current knowledge by utilizing the network weight matrix in short-term traffic forecasting. The network weight matrix captures both competitive and complementary links. We hypothesize this matrix represents a more realistic spatial dependence between links in traffic networks, and as a result boosts the accuracy of forecasting. To corroborate this hypothesis, we select a grid-like sub-network from the Minneapolis - St. Paul highway system comprising 140 traffic links and their traffic flow in 30-second increments for 2015. This topology discriminates the current research from previous studies, which typically picks a simple graph comprising only upstream and downstream links.

We structure the remainder of the research as follows. First, we review previous research adding spatial components to forecasting. Second, we discuss the importance of data resolution and prediction horizon in short-term traffic forecasting; particularly when the information of spatially correlated links is considered in the modeling. Third, we explain the data used for analysis and the study area. Fourth, we represent the three-dimensional data detrending method to detect complementary and competitive links and create the network weight matrix in a real-world example. Fifth, we represent a two-step method to identify the best look-back time window, and we validate the capability of network weight matrix using the 1<sup>st</sup>-order neighboring technique. We conclude the paper with summarizing key findings and suggesting follow-on research.

## 2 Revisiting Short-term Traffic Forecasting

What distinguishes short-term traffic forecasting from prevailing transport planning forecasts is the “resolution” and “horizon” of prediction. Resolution refers to the time interval used for traffic analysis, and horizon indicates the extent of time ahead in traffic forecasting. For instance, a model uses data with 30-second resolution to predict traffic conditions 10 minutes ahead. In the literature of short-term traffic forecasting, the data resolution and forecasting horizon typically ranges from 20 seconds (Dia, 2001) to 60 minutes (Whitlock and Queen, 2000).

Short-term traffic forecasting models offer three kinds of services: (1) high resolution, (2) long horizon, and (3) accurate prediction. However, you are able to pick only two. Using high resolution data to predict a long horizon would not be accurate. An accurate prediction with long horizon necessitates low resolution. An accurate prediction with high resolution necessitates short horizon. In other words, the higher the data resolution, the more difficult the prediction gets, and the larger the horizon, the less accurate the prediction becomes.

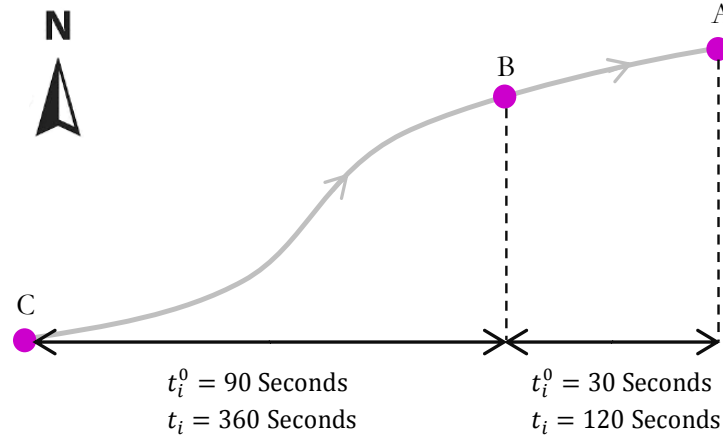
This section aims at delving into the literature of short-term traffic forecasting, and synthesizes studies from two aspects: (1) the resolution and horizon used in modeling, and (2) accuracy assessment of traffic forecasting using spatial components.

### 2.1 What is the Appropriate Data Resolution and Horizon?

Despite the key role of data resolution in short-term traffic forecasting, defining the appropriate resolution is mixed and crude. In 1993, Vythoulkas (1993) indicated 10-minute interval is the best

prediction interval, as the quality of information begins to decline in shorter intervals. The Highway Capacity [Manual \(2000\)](#) defined the 15-minute interval as an appropriate data resolution. In 2005, [Oh et al. \(2005\)](#) tested different aggregation from 1-minute to 10-minute intervals to pinpoint the most efficient data resolution for traffic prediction. They demonstrated that the prediction accuracy declines significantly for less than a 5-minute aggregation interval. A 4-minute to 6-minute aggregation interval is then suggested as desirable data resolution. In 2012, [Chen et al. \(2012\)](#) explored the appropriate aggregation interval among the 3-minute, 5-minute, 10-minute, and 15-minute data resolution. Using different traffic forecasting models, they concluded: “It is clear that the proposed prediction models (except the naïve models) are equivalent in performance under different aggregation time scales, when the residual time series are used in the prediction process.” However, digging down deeper into the result, it is found that the more aggregated the data, the more accurate prediction achieves. [Dunne and Ghosh \(2011\)](#) suggested that a 15-minute interval is the optimum forecasting resolution, and the accuracy of prediction is not significantly influenced beyond this interval. Notwithstanding differences, studies confirm that traffic flow exhibits strong fluctuations at shorter intervals, and thereby high resolution data diminishes the accuracy of prediction.

When spatial information is embedded in traffic forecasting models, the story is more intricate. This requires looking at the appropriate resolution problem from another angle. As alluded to previously, spatiotemporal approaches in short-term traffic forecasting benefit from the information of neighboring links. Using spatial information stems from the fact that many of the cars that will be on one link soon are already on the network upstream of the relevant location. A pertinent question is: how long does upstream traffic flow take to arrive downstream? Of course, the answer depends on the separation between two links, time-of-day, day-of-week, and traffic condition, which is defined as the dynamic travel time between two links. This travel time has a close relationship with the appropriate data resolution for maximizing the accuracy of prediction. Consider three sequential loop detectors in a highway corridor as [Figure 1](#).



**FIGURE 1 :** A toy example for data resolution analysis

In a free flow traffic regime, the travel time between detectors A and B is 30 seconds, and between A and C is 120 seconds. Thus, many of the cars that are detected by detectors B and C are observed at the location of detector A in 30 seconds and 120 seconds, respectively. So far

as spatiotemporal modeling is concerned, the past traffic flow of detectors A, B, and C is used to predict the traffic flow of detector A. In this case, the appropriate resolution for detector B is 30 seconds and for detector C is 120 seconds, and any other resolution misleads the accuracy of prediction. In a congested traffic regime, however, the lower resolution is needed to achieve an accurate prediction. In this case, the resolution for detector B is 120 seconds and for detector C is 480 seconds. Therefore, the appropriate resolution is highly correlated with travel time between the study link and its neighboring links, when the spatial information is included in traffic forecasting methods.

## 2.2 Does the Spatial Component Matter?

To assess the accuracy of short-term traffic forecasting models, studies employ various forms of error measurements. Table 1 summarizes prediction accuracy measurements.

**TABLE 1** : Summary of commonplace prediction accuracy measurements

Measurement	Abbreviation	Formula
Mean Relative Error	MRE	$\frac{1}{l} \sum_{i=1}^l \frac{ \hat{q}_i - q_i }{q_i}$
Mean Relative Percentage Error	MRPE	$\frac{1}{l} \sum_{i=1}^l \frac{ \hat{q}_i - q_i }{q_i} \times 100$
Mean Absolute Error	MAE	$\frac{1}{l} \sum_{i=1}^l  q_i - \hat{q}_i $
Mean Absolute Percentage Error	MAPE	$\frac{1}{l} \sum_{i=1}^l  q_i - \hat{q}_i  \times 100$
Mean Square Error	MSE	$\frac{1}{l} \sum_{i=1}^l (q_i - \hat{q}_i)^2$
Root Mean Square Error	RMSE	$\sqrt{\frac{1}{l} \sum_{i=1}^l (q_i - \hat{q}_i)^2}$
Mean Absolute Deviation	MAD	$\frac{1}{l} \sum_{i=1}^l  q_i - \bar{q}_i $
Max Absolute Error	MAE	$Max  q_i - \hat{q}_i $
Standard Deviation of the Absolute Error	SAE	$St.Dev.  q_i - \hat{q}_i $
Variance of Absolute Percentage Error	VAPE	$Var(\frac{ \hat{q}_i - q_i }{q_i}) \times 100$

Using MAPE, [Park et al. \(1998\)](#) examined the prediction accuracy of traffic flow in two sites with their upstream and downstream information. The results of Site 1 showed that using both upstream and downstream information in the backpropagation neural network model reduces the error by 30.2% in comparison with the model including downstream information. For Site 2, however, adding the upstream information increased the MAPE by 4.4%. This deficiency was justified by the presence of an exit ramp between upstream and downstream stations on Site 2.

In 2003, [Stathopoulos and Karlaftis \(2003\)](#) compared the accuracy prediction of univariate autoregressive integrated moving average (ARIMA) with a multivariate state-space model, which benefits from the information of upstream. The results demonstrated that adding the upstream information decreases the MAPE by 40.0%, although it varies in different traffic links. In a similar study, [Vlahogianni et al. \(2005\)](#) compared the univariate ARIMA with multivariate state-space model and indicated that adding the information of two upstream links diminishes the MRPE error by 18.7% in forecasting one step ahead.

A study compared ARIMA with both vector autoregression moving-average (VARMA) and space-time autoregressive integrated moving average (STARIMA) models, which include the

information of neighboring links ([Kamarianakis and Prastacos, 2003](#)). Using the RMSE measurement, they found the ARIMA models slightly outperformed the VARMA and STARIMA models. In total, the RMSE of the ARIMA models are about 5% less than the RMSE of both VARMA and STARIMA models.

In 2008, [Chandra and Al-Deek \(2008\)](#) predicted the travel speed in three stations with and without neighboring links. The results are mixed. They indicated adding the information of two downstream links reduces the RMSE and MAPE by 30.1% and 10.1%, respectively. Although adding the information of one upstream and two downstream diminishes the RMSE by 0.8%, it increases the MAPE by 5.2%. This study tested the other configurations such as one upstream and one downstream, two upstream and two downstream, and three upstream and one downstream. In general, the information of downstream was found more effective than upstream.

A more recent study ([Zeng and Zhang, 2013](#)) built an artificial neural network with and without upstream and downstream segments. The results of MAPE measurement underlined that incorporating either downstream or upstream segments decreases the prediction error by 8.9%. This diminution is 14.9% when the information of both segments are used for forecasting.

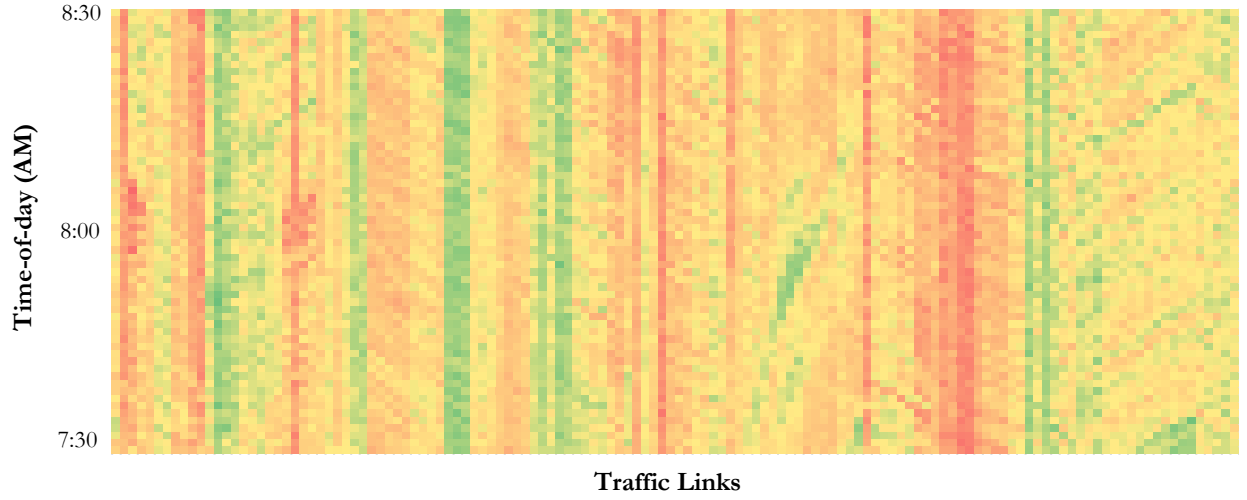
### 3 Data and Study Area

The core traffic flow dataset used in this research extracted from the Intelligent Roadway Information System (IRIS). IRIS is an open-source advanced traffic management system developed by the Minnesota Department of Transportation (MnDOT), which manages 5,452 vehicle detection stations, 433 ramp meters, 135 dynamic message signs, and 476 cameras installed in the Minneapolis - St. Paul highway system ([Darter et al., 2011](#)). This software records traffic flow of all traffic links in 30-second increments. For the purpose of analysis, we selected a sub-network of the Minneapolis - St. Paul highway system, located in the western suburbs as it has a grid-like topology, which includes both competitive and complementary links, and it consists of the busiest major highways in the Minneapolis - St. Paul highway. In particular, the selected sub-network is comprised of I-494, I-94, I-394, US 169, TH 212, TH 100, and TH 62 for the East-West and South-North directions. Our sample represents 140 traffic links, which their flow information is collected by 687 loop detectors. We obtained the traffic flow information for all Tuesdays of 2015 in two distinct times-of-day:

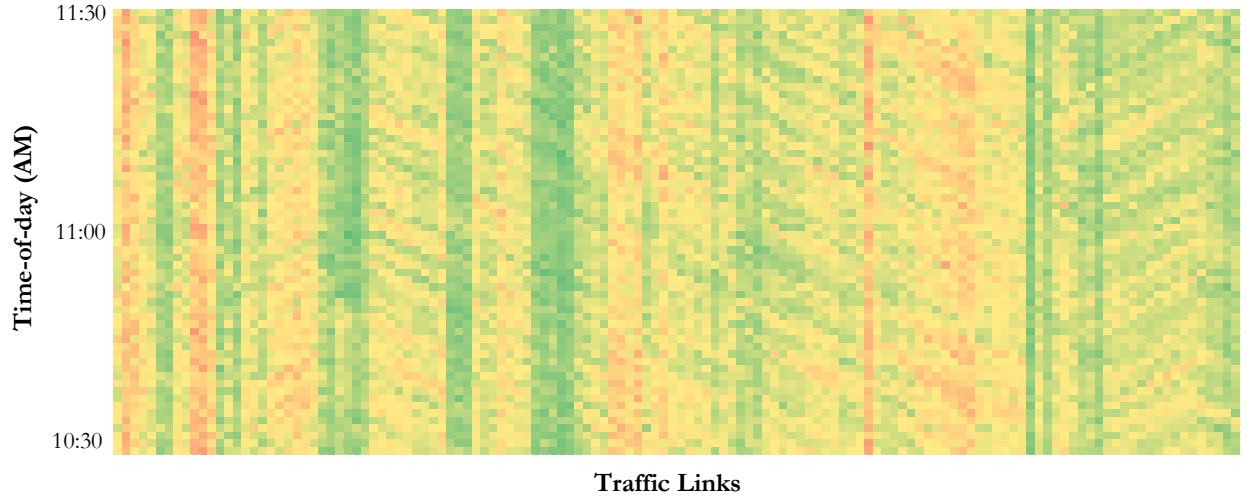
1. Morning rush hour: From 7:30-8:30 AM
2. Non-rush hour: From 10:30-11:30 AM

Tuesday is randomly selected as a weekday representative. We also randomly selected January 6<sup>th</sup> among Tuesdays of 2015 for validation, while controlling for two criteria: (1) This is not a holiday, and (2) There is no public event in the city to cause special traffic conditions. Figure 2 portrays the profile of traffic flow for all 140 links in both morning rush hour and non-rush hour time intervals. We use three color scales to depict the variation of traffic flow in each link over time of day. The green and red colors are chosen to show the first and last 10 percentile, respectively. As traffic flow increases in magnitude, the green spectrum approaches yellow, which indicates the 50 percentile of the values. With more increase in the value of traffic flow, the yellow spectrum becomes red. As shown, traffic flow is unevenly distributed over the network, in which some traffic

links carry a significant portion of demand. The higher density of red colors in the morning rush hour shows traffic flow in morning rush hour is significantly greater than non-rush hour.



(a) Morning rush hour



(b) Morning non-rush hour

**FIGURE 2** : Traffic flow of 140 links for Tuesday July 6<sup>th</sup>, 2015

## 4 Capturing Complementary and Competitive Links

At the link-level, traffic flow exhibits a temporal trend. This trend is a function of time-of-day and day-of-week, and has been the subject of interest in traffic forecasting. Traffic analysts tackled traffic flow forecasting, while benefiting from this temporal dependency of traffic conditions in a traffic link. At the network-level, traffic flow in a link of the network not only displays temporal correlation, but it also is spatially correlated with other links, although the magnitude of dependency varies from link to link. To observe the spatial correlation and detect the competitive and



complementary nature of traffic links, we need to appropriately eliminate temporal correlations in each link and between links. The data used in this research consists of every 1-minute block for two different one-hour time intervals from Tuesdays of 2015. The traffic flow then has a temporal trend in each one-hour time intervals and over Tuesdays of 2015. Although eliminating these two trends is necessary, they are not enough to remove temporal correlations between traffic links. Traffic flow of each link in a specific time intervals exhibits a temporal correlation with total flow existing in the network. Removing this correlation, which has been overlooked in traffic analyses, leaves the spatial correlation between traffic links (Ermagun et al., 2017). This section unpacks the statistical steps behind this three-dimensional data detrending.

We begin with Tuesdays of 2015 and two distinct times-of-day: (1) 7:30-8:30 AM and (2) 10:30-11:30 AM after controlling for all possible malfunctions in the real-world data. The algorithm, however, tests for all days of the week at different levels of data aggregation to ensure its strength and reliability in testing data. The output is a vector of traffic flow observations for each link and each day of the week, including  $m$  number of aggregated data points. In our example,  $m = 60$  as our data is smoothed over 1-minute for each one-hour time intervals. The vector of traffic flow is formulated as  $\mathbf{Y}(s, t) = (Y_{s,t,1}, \dots, Y_{s,t,m})$  for each link  $s$  and each day  $t$ . The notations  $s \in \{s_1, \dots, s_S\}$  and  $t = 1, 2, \dots, T$  represent traffic links and days of week, respectively. In our example,  $T = 52$  as our data is comprised of 52 consecutive weeks in 2015. We fit a robust location estimator to the data vectors  $\mathbf{Y}(s, t)$  for each link  $s \in \{s_1, \dots, s_S\}$ . This is captured by obtaining the minimizer  $\hat{\mu}(s) \in \mathbf{R}^m$  as per Equation 1. In this equation,  $\|\cdot\|_1$  is the  $L_1$ -norm of a vector. The output is the vector of medians for each location.

$$\sum_{t=1}^T \|\mathbf{Y}(s, t) - \mu(s)\|_1 \quad (1)$$

This helps us to remove the secular trend for each coordinate of the vector. To show the temporal correlation between consecutive time aggregation intervals, we apply an autoregression model to the detrended data obtained from the previous step. The best available autoregression model is selected testing lags of order 0 to 5, in which lags 0 and 5 signify no and high temporal correlation. We define  $\tilde{Y}(s, t, k) = Y(s, t, k) - \hat{\mu}(s, k)$  and fit the best available autoregression model under penalization the following model using the assumption that for each  $s$  and  $t$ , the sequence  $\{\varepsilon(s, t, k)\}$  is a mean zero, finite variance white noise sequence.

$$\tilde{Y}(s, t, k) = \sum_{j=0}^J \phi_{s,j} \tilde{Y}(s, t, k-j) + \varepsilon(s, t, k). \quad (2)$$

The second order stationarity is assumed for fitting the best autoregression model. The residuals of the model are then extracted for further analysis. Let  $S$  denotes the set of all links under consideration. The proportional residuals are defined in Equation 3 using residuals  $R(s, t, k)$  for each fixed  $t$  and  $k$ .

$$\tilde{R}(s, t, k) = \left[ \sum_{s \in S} R(s, t, k) \right]^{-1} R(s, t, k). \quad (3)$$

As a result, we derive Equation 4:

$$R(s, t, k) = Y(s, t, k) - \hat{\mu}(s, k) - \sum_{j=1}^J \hat{\phi}_{s,j}(Y(s, t, k-j) - \hat{\mu}(s, k-j)) \quad (4)$$

Following detection and elimination of trend and temporal correlations, we gather the spatial correlations patterns using the  $R(s, t, k)$  values. The first step is to elicit the adjacent correlation relations. We hence first obtain the adjacent correlation relations by finding serial correlations across each pair of link  $s_1$  and  $s_2$  for each time  $t$ . The output is formulated as follows:

$$C(s_1, s_2, t) = \text{Cor}(R(s_1, t, k), R(s_2, t, k)) \quad (5)$$

A strong yearly summary of these correlations is developed by taking the median  $\hat{C}_1(s_1, s_2)$  of  $\{C(s_1, s_2, 1), \dots, C(s_1, s_2, T)\}$ . If  $\hat{C}_1(s_1, s_2)$  is above a threshold  $c_1$ , we consider that links  $s_1$  and  $s_2$  are spatially correlated. In our example, we adopt  $c_1 = 0.10$ . Although we calculated both Kendall's and Spearman's correlation coefficients, we decided to base our correlation analysis on the Pearson correlation coefficient as we did not find any significant differences between the different correlation coefficient calculators.

Since we have now gathered the correlation structures, we then look into examining a longer range of complementary relations between links. We first compute the proportion of trend and temporal dependency adjusted residuals for each day  $t$  and each link  $s$ . This identifies the proportion of traffic flowing through station  $s$  on day  $t$  at each time aggregation step. Let these proportional residuals be  $\tilde{R}(s, t, k)$ . Across each pair of links  $s_1$  and  $s_2$  for each time  $t$ , we calculate the correlation according to Equation 6:

$$\tilde{C}(s_1, s_2, t) = \text{Cor}(\tilde{R}(s_1, t, k), \tilde{R}(s_2, t, k)). \quad (6)$$

A strong yearly summary of these is similarly developed by taking the median  $\hat{C}_2(s_1, s_2)$  of  $\{\tilde{C}(s_1, s_2, 1), \dots, \tilde{C}(s_1, s_2, T)\}$ , and obtain a negative or positive relation between stations  $s_1$  and  $s_2$  if  $\hat{C}_2(s_1, s_2) < -c_2$  for a chosen threshold  $c_2$ . In our example, we consider  $c_2 = 0.10$ .

We emphasize that for any pair of links  $s_1$  and  $s_2$ , if  $\hat{C}(s_1, s_2) > c_1$ , we consider these to be spatially correlated and, therefore, do not search for complementary relations. We consider if  $\hat{C}_2(s_1, s_2) < -c_2$ , for links pairs that are not spatially correlated and determine that there is a complementary relation between the two stations if this should happen.

Our calculations have been verified with other choices of thresholds and other tuning parameters of our algorithm. Several relations between links are gathered at very low levels of the thresholds  $c_1$  and  $c_2$ . Thresholds in the range 0.05 - 0.25 seem to be unaffected by algorithmic choices and produce consistent results. However, as the thresholds increase, the connections between the links gradually become weaker and become almost nonexistent past the 0.6 point. The sensitivity to  $c_2$  is higher than that to  $c_1$ . From these findings, we can reasonably determine that a very low threshold may produce relations that are false and temporary. On the other hand, a very high threshold could show relations between stations that are not genuine and consistent. Therefore, the choice of  $c_1 = c_2 = 0.1$  as a threshold presents a situation where the findings are consistent and there is a smaller chance of yielding results that show false relations, while not being too high that genuine relations are not found.

## 5 Capturing the best look-back time window

We broadly discussed the challenge of data resolution in traffic forecasting analysis, when the spatial component is utilized to improve the accuracy of prediction. To overcome this challenge, a naïve algorithm is developed to choose the best look-back time window corresponding to each neighboring ring. This algorithm finds the best look-back time window for all traffic links in two steps:

**Step 1:** For a selected time-of-day, the algorithm tests every  $t$ -second look-back time window in a specific exploring time interval in the model. Two main criteria are acknowledged to choose the  $t$ -second look-back time window. First, it should ensure the information of the immediate adjacent link is accurately collected when the travel time between two adjacent links is low. Second, the disaggregated data should be available for this time window. The selection of the time interval, however, should guarantee that the information of the farthest traffic link, which is significantly correlated with the study link, is accurately collected for forecasting. The value of both the  $t$ -second look-back time window and the exploring time interval depends on the physics of the study network, configuration of detectors, and the travel time gap between detectors. More precisely, the shortest travel time between the study link and its immediate adjacent link defines the  $t$ -second look-back time window, while the exploring time interval is measured by the largest travel time between the study link and its farthest traffic link in the study network.

In this study, a 30-second look-back time window and an one-hour exploring time interval are selected for two primary reasons: (1) Although the shortest travel time between some detectors is less than 30 seconds, the finest available data is every 30-seconds, and (2) in the extracted sub-network, the longest path starts from the intersection of I-494 with TH 100 and ends in the intersection of I-494 with I-94. The typical travel time of this path is around 25 minutes and 55 minutes in the uncongested and congested traffic conditions, respectively.

**Step 2:** The best look-back time window for each neighboring ring is selected according to the student's  $t$ -test statistic. The more accurate look-back time window for each neighboring ring, the higher fit of the  $t$ -test becomes.

To build the model and measure the student's  $t$ -test statistic, Equation 7 is adopted where the traffic flow of the current situation is modeled by the past traffic flow information of both the study link and the other 139 traffic links. In this equation,  $Q_t$  is a vector of traffic flow at time  $t$ ,  $Q_{t-0.5}$  is a vector of traffic flow 30 seconds earlier than time  $t$ , and  $\mathcal{W}$  represents the network weight matrix, which is the same as the correlation matrix ( $\hat{C}$ ) extracted in the preceding section. In order to compare  $\mathcal{W}$  with  $W$ , the  $w_{i,i}$  components of the network weight matrix are fixed to zero, as the spatial weight matrices typically cannot capture the self-influence of traffic link  $i$  upon itself.

$$Q_t = \alpha + \xi_{t-0.5} Q_{t-0.5} + \phi_{t-0.5} \mathcal{W} Q_{t-0.5} + \varepsilon_t \quad (7)$$

In the look-back time window selection, the following hypotheses are proposed:

**Hypothesis 1:** For upstream links, the farther the link, the larger look-back time window is needed.

**Hypothesis 2:** The look-back time window varies by traffic link, time-of-day, and day-of-week. In general, the look-back time window in congested regimes is larger than uncongested regimes.

To test these hypotheses, the best look-back time window is searched for at 8:30 AM and

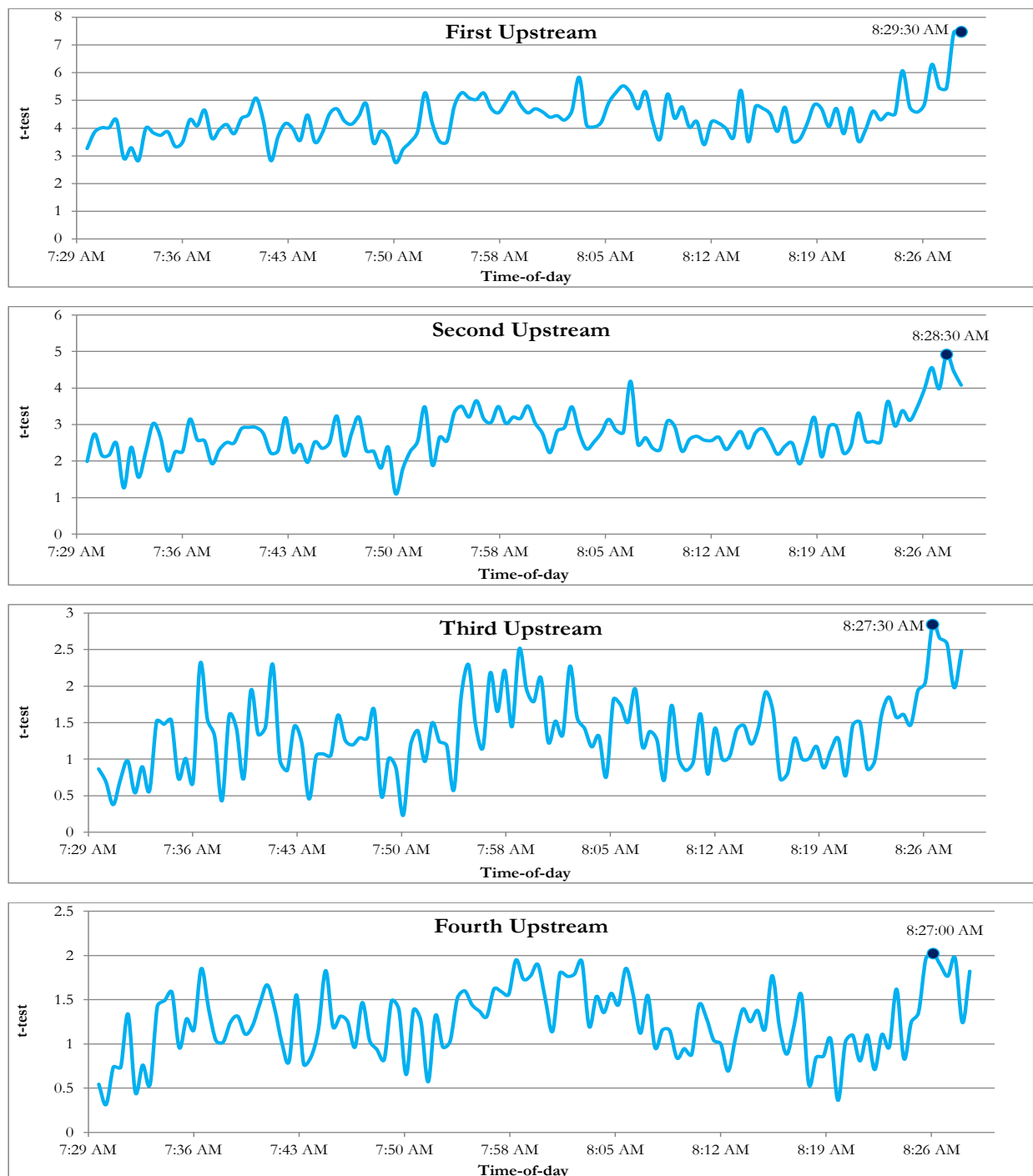
11:30 AM on January 6<sup>th</sup> 2015. This date is randomly selected for the sake of analysis among Tuesdays of 2015. To validate the results, the proposed two-step algorithm is applied on all traffic links for four upstream links. The student's t-test statistic results are depicted in Figure 3 for 8:30 AM and in Figure 4 for 11:30 AM. Figure 3 demonstrates that the best look-back time window is larger for the farther traffic links. For instance, this time window is 8:29:30 AM for the first upstream and 8:27:00 AM for the fourth upstream. A point of emphasis is that this look-back time window is an average over all traffic links. In the extracted sub-network from the Minneapolis - St. Paul highway system, the configuration of detectors causes an average of 30 seconds travel time between the two serial adjacent traffic links in a congested regime; while, each upstream link has a unique look-back time window that is a function of its travel time from downstream link. Looking at Figure 3, it is found that the strength of correlation between the link in question and its upstream significantly diminishes with distance. This is worth conducting further research to test the applicability of the first "law" of geography among the upstream links.

The results highlight that the look-back time window for the uncongested traffic regime is shorter than the congested one. Comparing the strength of correlation between the uncongested and congested traffic regimes, it is understood that the 30-second look-back time window does not perfectly catch the information of the upstream. The lack of more disaggregated data renders it almost impossible to search the best look-back time window, as we assume this should be less than 30 seconds in uncongested traffic regimes due to the average distance between detectors. In justification of this assumption, the distance between each link and its four upstream links is calculated for 140 traffic links of the extracted sub-network. Table 2 summarizes the statistic of this calculation.

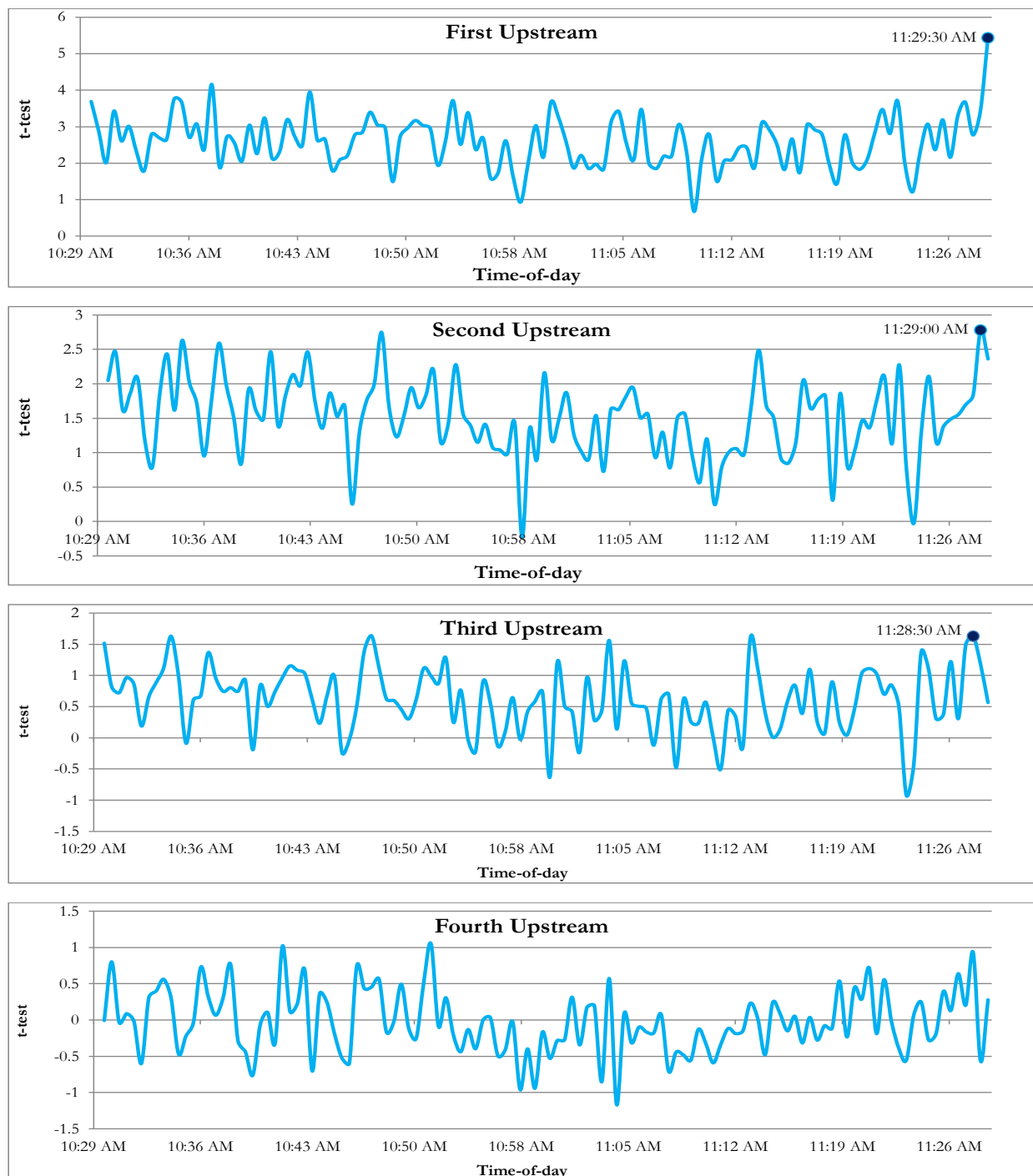
As shown in Table 2, the distance between each link and its immediate upstream link fluctuates between 0.01 and 3.54 Km with an average of 0.9 Km. Knowing 96.5 Km/hr (60 mph) as the speed limit in the Minneapolis - St. Paul highways, and considering that drivers typically drive 8-16 Km/hr (5-10 mph) over the speed limit in uncongested conditions, it is plausible to assume the distance between two detectors is traveled in less than 30 seconds. It is also inferred from Table 2 that the standard deviation of distance between traffic links and their upstream escalates with an increase in the order of neighboring links. For example, the standard deviation for the third and fourth upstream is 1.9 and 2.3 times the first upstream. This indicates the distance between each traffic link and its upstream is spread out over a wider range of values by an increase in the order of neighboring links. It explains why the strength of correlation between the study link and its upstream declines in the higher order neighbors as shown in Figure 3 and Figure 4. This leads to the conclusion that holding an identical look-back time window for all traffic links misrepresents the results, particularly in the higher order neighbors.

**TABLE 2** : The statistics of distance from four upstream links over 140 traffic links

Upstream	Distance statistics (Kilometer)			
	<i>Min</i>	<i>Max</i>	<i>Average</i>	<i>St. Dev.</i>
<b>First</b>	0.01	3.54	0.90	0.48
<b>Second</b>	0.90	4.51	1.92	0.74
<b>Third</b>	1.46	5.57	2.86	0.94
<b>Fourth</b>	2.00	6.75	3.75	1.11



**FIGURE 3 :** Look-back time window of upstream links for 8:30 AM, January 6<sup>th</sup> 2015



**FIGURE 4 :** Look-back time window of upstream links for 11:30 AM, January 6<sup>th</sup> 2015

## 6 Traffic Flow Forecasting: First-order Neighboring Approach

This section discusses the results of modeling, where the 1<sup>st</sup>-order neighboring links are selected as potential spatial information. This aims not only to validate the network weight matrix approach, but also to examine the capability of the network econometrics model against the traditional approaches. Three distinct models are developed for different times-of-day for January 6<sup>th</sup> 2015. It enables us to scrutinize the role of weight matrices in different traffic regimes.

**Model 1:** This model includes the 30-second look-back time window of the study link as Equation 8. In this equation,  $Q_t$  is a vector of traffic flow at time  $t$ ,  $Q_{t-0.5}$  is a vector of traffic flow 30 seconds earlier than time  $t$ .

$$Q_t = \alpha + \xi_{t-0.5}Q_{t-0.5} + \varepsilon_t \quad (8)$$

**Model 2:** This model includes both the 30-second look-back time window of the study link and the 1<sup>st</sup>-order neighboring links. In this model, the spatial interaction is captured by the traditional spatial weight matrix, which comprises binary components. In the 1<sup>st</sup>-order spatial weight matrix, the one elements signify two links directly connected by a node. Equation 9 shows the model. In this equation,  $W$  stands for the spatial weight matrix.

$$Q_t = \alpha + \xi_{t-0.5}Q_{t-0.5} + \varphi_{t-0.5}WQ_{t-0.5} + \varepsilon_t \quad (9)$$

**Model 3:** Like Model 2, this model includes both the 30-second look-back time window of the study link and 1<sup>st</sup>-order-neighboring links. However, the spatial interaction is captured by the network weight matrix. Equation 7 demonstrates this model.

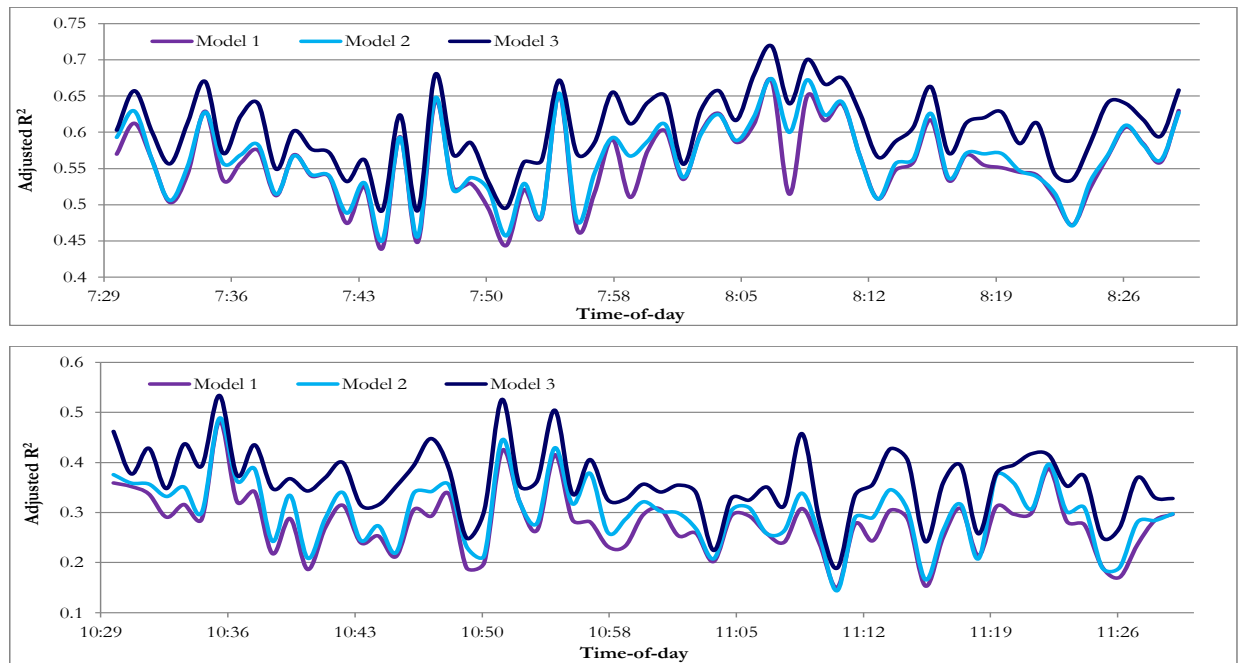
### 6.1 Network Weight Matrix Validation

To test the network weight matrix, Models 1, 2, and 3 are developed for every minute in both the morning rush hour and non-rush hour of Tuesday, January 6<sup>th</sup>, 2015. Thereupon, 360 models were developed to test the network weight matrix in a real-world example. Unlike the previous studies that typically select a few links for the validation and prediction accuracy, this study develops the models on all 140 traffic links extracted from the Minneapolis - St. Paul highway system. Thus, traffic links represent the observations in the modeling process.

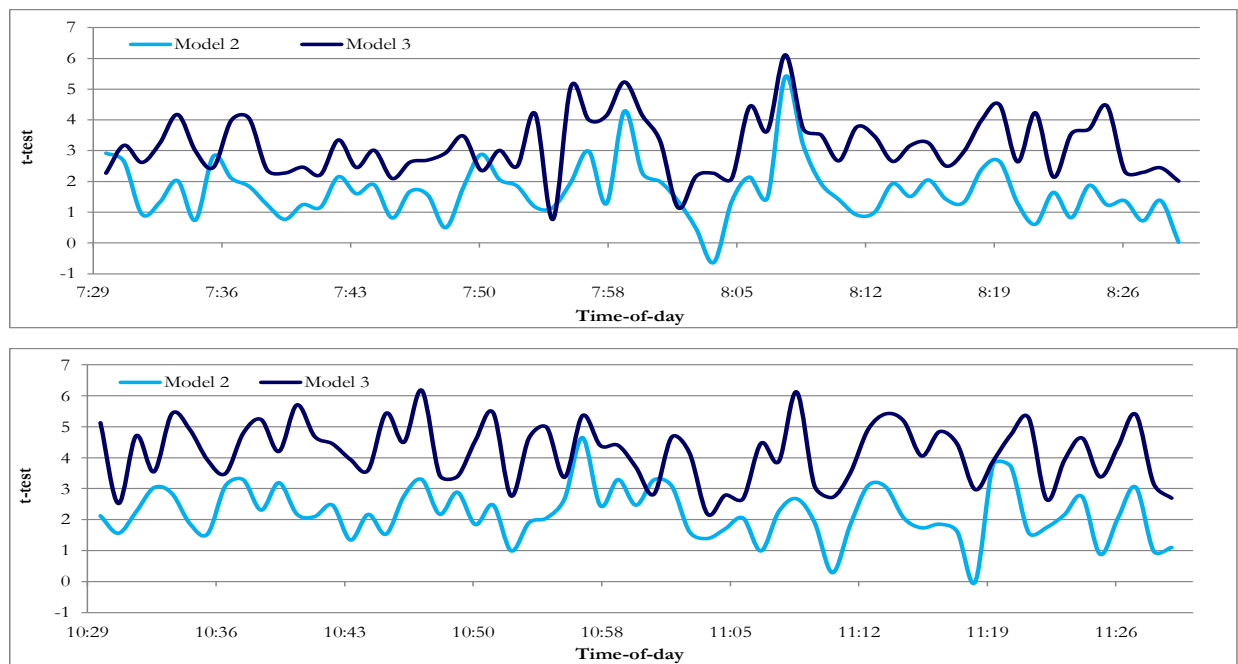
The Adjusted  $R^2$  and student's t-test for each model are depicted in Figures 5 and 6, respectively. In all models, for simplicity, the 30-second look-back time window is applied for all the 1<sup>st</sup>-order neighboring links. This restricted assumption largely reduces the accuracy of models, however it is applicable in this case for two reasons: (1) the models aim to compare the traditional spatial weight matrix with the proposed network weight matrix, and validate the robustness of the network weight matrix, and (2) the 30-second look-back time window has a strong correlation in the 1<sup>st</sup>-order neighboring links, as the standard deviation of the distance between traffic links and their first upstream is low and the values tend to be close to the average in our sample.

Pertaining to the Adjusted  $R^2$  measurement, Model 2 performs fairly close to Model 1 in both rush hour and non-rush hour conditions. In rush hour, the Adjusted  $R^2$  of Model 1 fluctuates between 0.43 and 0.67 with an average of 0.55. The Adjusted  $R^2$  of Model 2 fluctuates between 0.45 and 0.67 with an average of 0.56 over time. Likewise, the performance of both models is close in non-rush hour conditions, although Model 2 slightly outperforms Model 1. As shown in Figure





**FIGURE 5 :** Adjusted  $R^2$  of 360 models, January 6<sup>th</sup> 2015



**FIGURE 6 :** Student's t-test of 1<sup>st</sup>-order neighboring dependency, January 6<sup>th</sup> 2015



5, the general fit of the models significantly drops during the non-rush hour period for two reasons. First, the 30-second look-back time window is reasonably high for the non-rush hour in light of the detector configuration in the case of this study. In the non-rush hour, the traffic network typically experiences free-flow speed, and vehicles pass two consecutive loop detectors in a fairly low time in comparison with rush hour period. One may criticize and question the models due to their low general fit. However, it should be kept in mind that this study does not aim to represent the best model, but targets to validate the network weight matrix and to compare the three models. In non-rush hour, the Adjusted  $R^2$  of Model 1 ranges from 0.14 to 0.48 with an average of 0.27. Similarly, the Adjusted  $R^2$  of Model 2 ranges from 0.14 to 0.48; however, the average is slightly more than Model 1 and equals 0.30. Looking at Figure 5, it is identified that Model 3 outperforms Model 1 and Model 2 in both rush and non-rush hours. As expected, this gain in performance is higher in the non-rush hour regime, where the system is less complex. The Adjusted  $R^2$  of Model 3 ranges from 0.49 to 0.71 with an average of 0.60 in rush hour and from 0.18 to 0.53 with an average of 0.36 in non-rush hour. This indicates, on average, Model 3 outperforms Model 1 and Model 2 by 9% and 7% in rush hour traffic conditions, respectively. In non-rush hour traffic conditions, Model 3 outperforms Model 1 and Model 2 by 33% and 20%, respectively.

Pertaining to the t-statistic test, the results demonstrate that the effect of the spatial weight matrix is not stable and valid over time. This is in line with previous studies (Ermagun and Levinson, 2016c), which tested the robustness of the spatial weight matrix in a simulation-based environment. In the morning rush hour, the coefficient of the spatial weight matrix is insignificant at the 95% confidence interval in 57.3% of the cases. This number is 29.5% in the morning non-rush hour traffic time. However, the coefficient of the network weight matrix is statistically significant at the 95% confidence interval in all cases. This leads to the conclusion that the traditional spatial weight matrix is unable to capture the accurate spatial dependency between traffic links, particularly in a congested condition.

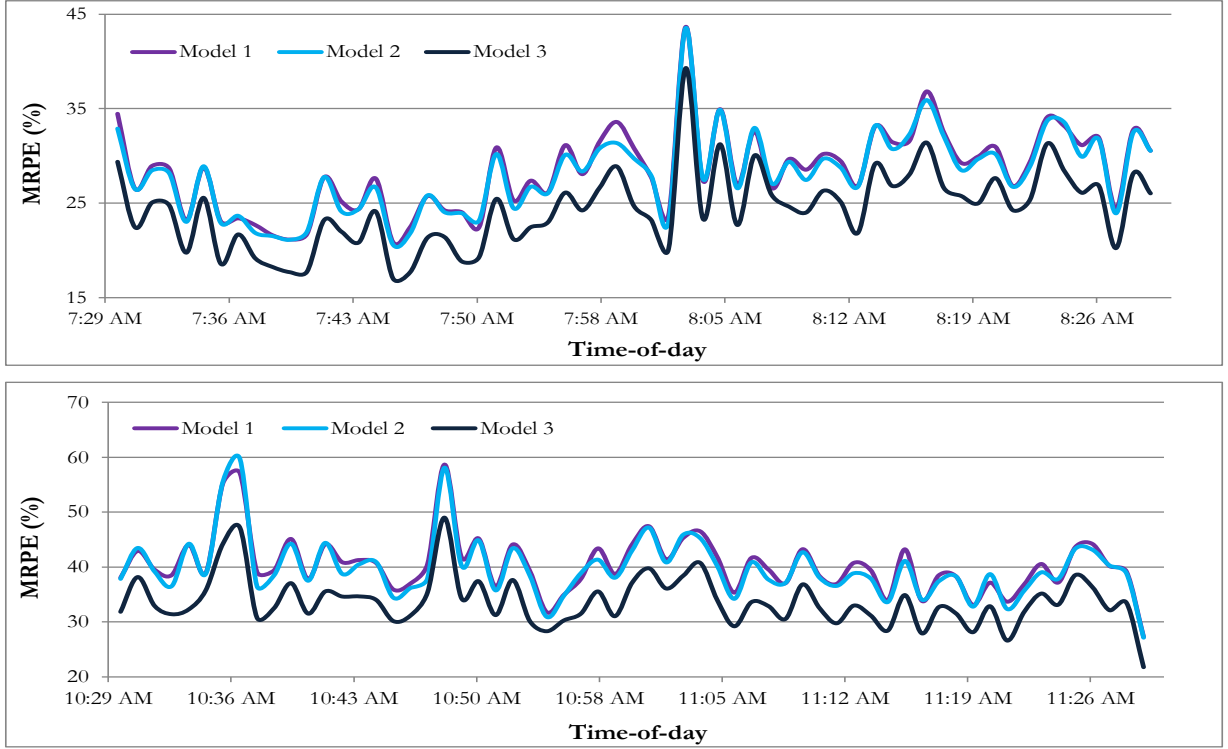
## 6.2 Prediction Accuracy

To cross validate the developed models and the network weight matrix, the 30-second traffic flow of the first Tuesday of 2016 (January 5<sup>th</sup>, 2016) is extracted for two different times of day:

1. Morning rush hour: From 7:30-8:30 AM
2. Non-rush hour: From 10:30-11:30 AM

This day is selected as the models were developed on the first Tuesday of 2015. The developed models are utilized to predict the one-step ahead traffic flow of each link. For example, the traffic flow at 7:59:30 AM is used to predict the traffic flow at 8:00:00 AM. Each of 360 models developed on January 6<sup>th</sup>, 2015, are applied at its exact time-of-day on January 5<sup>th</sup>, 2016. To measure the prediction accuracy of the models, the mean relative percentage error is calculated for each link. For every 1-minute, the mean relative percentage error of models is averaged over traffic links. Figure 7 shows the results of the error. As shown, the mean relative percentage error of Model 3 is significantly lower than Model 1 and Model 2 in all cases over time. In rush hour, the MRPE of Model 3 varies from 17.07% to 39.2% with an average of 24.3% over time. For Model 1 and Model 2, the average of MRPE over time is 28.3% and 28.0%, respectively. This indicates

that Model 3 outperforms Model 1 by 14.1% and Model 2 by 13.2%, on average. Although the prediction accuracy of the models in non-rush hour is lower than rush hour, Model 3 surpasses Model 1 by 16.7% and Model 2 by 15.3%, on average.



**FIGURE 7 :** Mean relative percentage error of 360 models, January 5<sup>th</sup> 2016

## 7 Summary and Conclusions

Despite its complexity, the short-term traffic is predictable with a deep understanding of relationships between traffic components. Under this conviction, this study employed network weight matrices to quantify the dependency between traffic links. We tested and validated the network weight matrix in a real-world example. Traffic flow of 140 highway traffic links in a sub-network of the Minneapolis - St. Paul freeway was modeled for a random Tuesday of 2015. To cross validate the developed models and the network weight matrix, the 30-second traffic flow of Tuesday, January 5<sup>th</sup>, 2016 was extracted for both rush hour and non-rush hour periods. The results of the cross-validation showed the network weight matrix outperforms the traditional way of capturing spatial dependence between traffic links. Aside from the superiority in forecasting, a remarkable capability of the network weight matrix was its stability and robustness over time, which was not witnessed in spatial weight matrix. In addition, this study proposed a naïve two-step algorithm to search and identify the best look-back time window for upstream links. The following items are the noteworthy findings:

- For the 1<sup>st</sup>-order neighboring analysis, the spatial weight matrix is unstable over time-of-day, while the network weight matrix is robust in all cases.

- For the 1<sup>st</sup>-order neighboring analysis, the performance of the network weight matrix in non-rush hour traffic regimes is significantly better than rush hour traffic regimes.
- The best look-back time window depends on the travel time between two study detectors, and it varies by time-of-day and traffic link.
- The best look-back time window is significantly shorter in uncongested traffic conditions than congested traffic conditions.
- The network weight matrix outperforms the traditional way of capturing spatial dependence between traffic links. Averaging over all traffic links and time, this superiority is about 13.2% in rush hour and 15.3% in non-rush hour, when only the 1<sup>st</sup>-order neighboring links are embedded in modeling.

Network weight matrices are able to enhance the accuracy of traffic prediction. Subsequently, the network weight matrix solution has the potential application in, but not limited to, ramp metering signal control, freeway control, and ITS applications. Although this study validated the network weight matrices for short term traffic forecasting, they have potential for implementation for other transport problems, and in geography, regional, and social network sciences for which links or nodes may be either competitive or complementary with each other. The following suggestions are made for further research:

First, to validate the capability of the network weight matrix, this research targeted at the 1<sup>st</sup>-order neighboring links in a real-world network problem, and overlooked the role of all competitive and complementary links in short-term traffic forecasting. Testing the potential of competitive links and embedding this information in forecasting methods have room to grow. If the traffic flow of a link is predicted by its competitive links, it not only enhances the prediction accuracy, but may substitute for adjacent links in the absence of their information.

Second, traffic networks possess distinct characteristics that change over time and space. For example, capacity changes spatially, but may or may not change temporally. The betweenness depends upon the measurement method. It may or may not change temporally but must change spatially. Total system demand (e.g. regional population) changes temporally, but is uniform spatially. This research modeled the traffic flow using a temporal lag. However, the proposed network econometrics approach is able to incorporate both spatial and temporal exogenous variable.

Third, although this research employed the network weight matrix with theoretical assertions for the short-term traffic flow forecasting, it is envisaged that the proposed network weight matrices could find a wide range of applications in traffic forecasting during predictable and unpredictable incidents. Hence, a research avenue is open for further validation in follow-up traffic forecasting studies.

## References

Chandra, S. and Al-Deek, H. Cross-correlation analysis and multivariate prediction of spatial time series of freeway traffic speeds. *Transportation Research Record: Journal of the Transportation Research Board*, (2061):64–76, 2008.

- Chen, C., Wang, Y., Li, L., Hu, J., and Zhang, Z. The retrieval of intra-day trend and its influence on traffic prediction. *Transportation research part C: emerging technologies*, 22:103–118, 2012.
- Darter, M., Swanston, T., Yen, K., Ravani, B., and Lasky, T. Technical report documentation page. 2011.
- Dia, H. An object-oriented neural network approach to short-term traffic forecasting. *European Journal of Operational Research*, 131(2):253–261, 2001.
- Dunne, S. and Ghosh, B. Regime-based short-term multivariate traffic condition forecasting algorithm. *Journal of Transportation Engineering*, 138(4):455–466, 2011.
- Ermagun, A. and Levinson, D. Spatiotemporal traffic forecasting: Review and proposed directions. 2016a.
- Ermagun, A. and Levinson, D. An introduction to the network weight matrix. *Geographical Analysis*, 2016b.
- Ermagun, A. and Levinson, D. M. Development and application of the network weight matrix to predict traffic flow for congested and uncongested conditions. 2016c.
- Ermagun, A., Chatterjee, S., and Levinson, D. Using temporal detrending to observe the spatial correlation of traffic. *PloS one*, 12(5):e0176853, 2017.
- Kamarianakis, Y. and Prastacos, P. Forecasting traffic flow conditions in an urban network: comparison of multivariate and univariate approaches. *Transportation Research Record: Journal of the Transportation Research Board*, (1857):74–84, 2003.
- Kamarianakis, Y., Prastacos, P., and Kotzinos, D. Bivariate traffic relations: A space-time modeling approach. *AGILE proceedings*, pages 465–474, 2004.
- Manual, H. C. Highway capacity manual. *Washington, DC*, 2000.
- Oh, C., Ritchie, S., and Oh, J.-S. Exploring the relationship between data aggregation and predictability to provide better predictive traffic information. *Transportation Research Record: Journal of the Transportation Research Board*, (1935):28–36, 2005.
- Okutani, I. and Stephanedes, Y. J. Dynamic prediction of traffic volume through kalman filtering theory. *Transportation Research Part B: Methodological*, 18(1):1–11, 1984.
- Park, B., Messer, C., and Urbanik II, T. Short-term freeway traffic volume forecasting using radial basis function neural network. *Transportation Research Record: Journal of the Transportation Research Board*, (1651):39–47, 1998.
- Smith, B. L., Williams, B. M., and Oswald, R. K. Comparison of parametric and nonparametric models for traffic flow forecasting. *Transportation Research Part C: Emerging Technologies*, 10(4):303–321, 2002.

- Stathopoulos, A. and Karlaftis, M. G. A multivariate state space approach for urban traffic flow modeling and prediction. *Transportation Research Part C: Emerging Technologies*, 11(2):121–135, 2003.
- Sun, S., Zhang, C., and Zhang, Y. Traffic flow forecasting using a spatio-temporal bayesian network predictor. In *International Conference on Artificial Neural Networks*, pages 273–278. Springer, 2005.
- Vlahogianni, E. I., Karlaftis, M. G., and Golias, J. C. Optimized and meta-optimized neural networks for short-term traffic flow prediction: a genetic approach. *Transportation Research Part C: Emerging Technologies*, 13(3):211–234, 2005.
- Vythoulkas, P. Alternative approaches to short term traffic forecasting for use in driver information systems. *Transportation and traffic theory*, 12:485–506, 1993.
- Whitlock, M. E. and Queen, C. M. Modelling a traffic network with missing data. *Journal of Forecasting*, 19(7):561–574, 2000.
- Zeng, X. and Zhang, Y. Development of recurrent neural network considering temporal-spatial input dynamics for freeway travel time modeling. *Computer-Aided Civil and Infrastructure Engineering*, 28(5):359–371, 2013.